PB86-194487

Use of a Graph Theoretic Similarity Index in Prediction Studies of Linear Discriminants and Models

Pennsylvania State Univ.

Prepared for

Health Effects Research Lab.
Research Triangle Park, NC

May 86

U.S. Department of Commerce
National Technical Information Service
NTIS.

PB86-194487

EPA/600/D-86/099
May 1986

USE OF A GRAPH THEORETIC SIMILARITY INDEX IN PREDICTION STUDIES OF LINEAR DISCRIMINANTS AND MODELS

by

Edward P. Jaeger and Peter C. Jurs

The Pennsylvania State University
University Park, PA 16802

EPA Grant CR811037

EPA Project Officer
James R. Rabinowitz

HEALTH EFFECTS RESEARCH LABORATORY
OFFICE OF RESEARCH AND DEVELOPMENT
U.S. ENVIRONMENTAL PROTECTION AGENCY
RESEARCH TRIANGEL PARK, NC 27711

REPRODUCED BY
NATIONAL TECHNICAL
INFORMATION SERVICE
U.S. DEPARTMENT OF COMMERCE
SPRINGFIELD, VA. 22161

TECHNICAL REPORT DATA
(Please read Instructions on the reverse before completing)

1. REPORT NO. EPA/600/D-86/099	2.	3. RECIPIENT'S ACCESSION NO.
4. TITLE AND SUBTITLE USE OF A GRAPH THEORETIC SIMILARITY INDEX IN PREDICTION STUDIES OF LINEAR DISCRIMINANTS AND MODELS		5. REPORT DATE May 1986 6. PERFORMING ORGANIZATION CODE
7. AUTHOR(S) Edward P. Jaeger and Peter C. Jurs		8. PERFORMING ORGANIZATION REPORT NO.
9. PERFORMING ORGANIZATION NAME AND ADDRESS Pennsylvania State University, University Park, PA 16802		10. PROGRAM ELEMENT NO. 11. CONTRACT/GRANT NO. CR811037
12. SPONSORING AGENCY NAME AND ADDRESS Health Effects Research Laboratory - RTP, NC Office of Research and Development US Environmental Protection Agency Research Triangle Park, NC 27711		13. TYPE OF REPORT AND PERIOD COVERED 14. SPONSORING AGENCY CODE EPA/600/11
15. SUPPLEMENTARY NOTES		

16. ABSTRACT

A goal of many structure-activity and structure-property studies is to develop the capability to predict the activity or property of interest for previously untested compounds. A quantitative model or a discriminant is developed from a training set of molecules with known activities or properties. The best test of the predictive ability of a model or a discriminant is to predict the activities or properties of a set of unknown compounds, so the agreement between predicted and observed values can be judged. Unknown compounds which are structurally similar to those of the training set stand the best chance of being predicted accurately. Structural similarity can be assessed with atom pairs. The metric is based on the relative frequencies of all unique atom pairs in each compound being compared. The method can be used to compare two molecules, or it can be used to characterize a set of compounds, or it can be used to compare a few compounds (a prediction set) to a large collection of compounds (a training set). The characteristics of the atom pair similarity measure are discussed and examples of its use in the selection of prediction set members in several structure-activity and structure-property studies are shown.

17. KEY WORDS AND DOCUMENT ANALYSIS

a. DESCRIPTORS	b. IDENTIFIERS/OPEN ENDED TERMS	c. COSATI Field/Group

18. DISTRIBUTION STATEMENT RELEASE TO PUBLIC	19. SECURITY CLASS (This Report) UNCLASSIFIED	21. NO. OF PAGES 12
	20. SECURITY CLASS (This page) UNCLASSIFIED	22. PRICE

EPA Form 2220-1 (Rev. 4-77) PREVIOUS EDITION IS OBSOLETE

NOTICE

This document has been reviewed in accordance with U.S. Environmental Protection Agency policy and approved for publication. Mention of trade names or commercial products does not constitute endorsement or recommendation for use.

USE OF A GRAPH THEORETIC SIMILARITY INDEX IN PREDICTION STUDIES OF LINEAR DISCRIMINANTS AND MODELS

Edward P. Jaeger and Peter C. Jurs

152 Davey Laboratory, Chemistry Department, The Pennsylvania State University, University Park, PA 16802

ABSTRACT

A goal of many structure-activity and structure-property studies is to develop the capability to predict the activity or property of interest for previously untested compounds. A quantitative model or a discriminant is developed from a training set of molecules with known activities or properties. The best test of the predictive ability of a model or a discriminant is to predict the activities or properties of a set of unknown compounds, so the agreement between predicted and observed values can be judged. Unknown compounds which are structurally similar to those of the training set stand the best chance of being predicted accurately. Structural similarity can be assessed with atom pairs. The metric is based on the relative frequencies of all unique atom pairs in each compound being compared. The method can be used to compare two molecules, or it can be used to characterize a set of compounds, or it can be used to compare a few compounds (a prediction set) to a large collection of compounds (a training set). The characteristics of the atom pair similarity measure are discussed and examples of its use in the selection of prediction set members in several structure-activity and structure-property studies are shown.

INTRODUCTION

Multivariate analysis is one of several computer-assisted methods of studying structure-activity relationships (SAR). In a multivariate SAR study, a correlation is sought between numerical descriptions of molecules and some measure of their biological activity or physical property. The compounds are represented by points in a high-dimensional space, where each dimension corresponds to a molecular structure descriptor. Multivariate structure-activity studies have two general purposes: the prediction of activity for previously untested compounds and the potential explanation of some mechanistic questions in the activity.

A concern is the quality of the prediction results. The models or discriminants are generated using a limited number of compounds (a training set) and can not completely describe the universe of organic compounds. Therefore, it is important to determine whether the prediction of the activity of a compound of unknown activity is an interpolation or an extrapolation. The prediction will be an interpolation if the high-dimensional point representing the test compound falls within the cluster of the training compounds in the data space. The prediction will be an extrapolation if the point representing the test compound falls outside those used to develop the model or discriminant. Using standard statistical methods it is possible to look for outliers within the data space of the molecular descriptors. If the molecular description of a compound is found to

be an outlier, the prediction results for that compound may be in error. Sometimes the molecular descriptors selected to form the model will be focused on a particular region of the molecules in the training set. As long as the moiety of interest forms the active portion of a prediction compound the prediction results are likely to be valid. It is possible that the described moiety could be just a small part of a prediction compound or it could be that there are other parts of the compound more related to the overall activity. Thus, it is important to assess the overall similarity of a test molecule compared to the training set compounds before accepting the prediction results.

The comparison of the overall similarity of molecules whose activities are to be predicted to the compounds in the training set has been hampered by the lack of a suitable way of describing entire molecules and comparing them. Carhart, et al. recently reported the use of small structural features known as atom pairs to describe an entire molecule and to compare it to other molecules (1). We report the use of the atom pair similarity metric and extensions for the comparison of two molecules, the characterization of a set of compounds, and the comparison of a few compounds to a large collection of compounds.

DESCRIPTION OF ATOM PAIRS

The atom pairs method of Carhart is a molecular description based on topology. A molecule to be coded is broken into every possible pair of atoms. Each atom is described by its elemental type, the number of non-hydrogen attachments, and the number of electrons involved in pi bonding. A complete atom pair description includes an atomic description for each atom and a count of the number of bonds separating the two atoms. A list of all unique atom pairs and a frequency count serves as the description for the whole molecule. The atom pairs description encodes only topological information. It is not able to distinguish between stereoisomers or some positional isomers.

The atomic environment of the carbon in carbon monoxide is described by notation for a C atom, 1 non-hydrogen neighbor, and 2 pi electrons. A similar description for the oxygen atom of carbon monoxide is O atom, 1 non-hydrogen neighbor, and 2 pi bonding electrons. The atom pair description for a complete molecule is a list of all unique atom pairs and their frequencies. For carbon monoxide the list consists of just one atom pair:

C:X--2--O:X

Colons and periods are used to indicate one or two pi bonding electrons. The X's represent non-hydrogen neighbors, and the numeral represents the number of nodes passed when traveling from one atom to the next in the molecular graph and is therefore the number of bonds separating the two environments plus one. The atom pair description for benzene consists of three atom pairs:

C.X2--4--c.X2
C.X2--3--C.X2
C.X2--2--C.X2

with frequencies of 3, 6 and 6 respectively.

SIMILARITY CALCULATIONS

The topological similarity of two molecules, each described by a list of unique atom pairs, is obtained by comparing the contents of each atom pair list. If an atom pair is found in both lists, it represents a ~~common~~ feature common to the two molecules. The similarity index is formed by counting all the common occurrences of all atom pairs and dividing by the sum of all atom pairs, common and noncommon, present in both molecules. This full similarity index ranges from zero (no common atom pairs) to unity (all atom pairs are common). The full similarity index for benzene and toluene is 0.556. There are a total of 10 common atom pairs (three unique types) and 36 atom pairs in all (15 from benzene and 21 from toluene). The full similarity index of toluene and o-xylene is 0.699, and the similarity of o-xylene and m-xylene is 0.679.

We have developed several additional indices based on atom pairs to handle specific circumstances. All the indices involve a ratio of the number of common atom pairs to the total number of atom pairs, but the weight given to each specific atom pair varies. The full similarity index treats all atom pairs equally. The prototype atom pair similarity index permits the exclusion of atom pairs present in a specific prototype molecule from the calculation. If a data set of substituted naphthalenes were being studied, the similarity calculation could be focused on the substituents by subtracting the atom pairs found in the prototype naphthalene from the calculations.

A third similarity index used to compare compounds within a data set is known as the frequency similarity index. As in the prototype similarity index, the intent of the frequency index is to deemphasize the effect of prevalent atom pairs on the calculation by weighting them less in the calculation. The weighting factor for each atom pair's contribution to the similarity equation is inversely proportional to the frequency with which it occurs in the entire set of compounds.

DATA SETS

In the work by Carhart, et al., the atom pair similarity measure was used to search large databases for structures similar to various test structures. We are concerned with its use for characterizing sets of compounds for homogeneity or heterogeneity and its use in justifying the exclusion or inclusion of test compounds in a prediction set. For this work we employed three sets of compounds which have been used for previous SAR studies. The N-nitrosamine (NNO) data set consists of 150 compounds with known carcinogenicity and was studied by Rose and Jurs (2). The 92 N-nitrosamines which form the prediction set were reported by Lijinsky (3) and Preussman and Stewart (4). The 213 9-anilino acridines (9-AA) were studied by Henry, Jurs, and Denny (5), and 49 additional compounds have been used here for a total of 262. The biological activity of concern for the 9-AAs is anti-tumor activity. The polycyclic aromatic compounds (PAC) data set consists of 447 compounds whose carcinogenicity was studied by Whalen-Pedersen and Jurs (6). The structures of each compound in the three data sets were entered onto disk files on the Chemistry Department's PRIME 750 computer using the ADAPT software system (7). Software was developed to generate atom pairs for each structure and to calculate similarities using the various metrics.

RESULTS AND DISCUSSION

The NNOs are a relatively heterogeneous mix of compounds with the central defining feature being the three atom N-nitroso functional group. The attachments to the NNO moiety vary from methyl groups through 14 member alkyl chains to saturated and aromatic ring systems. The atom pair similarities of this set provides a means of describing and quantifying this diversity. An average similarity was determined for each of the 150 members of the NNO set by calculating the similarity with every other member of the set. A histogram of these values is presented in Figure 1. The mean for the NNO

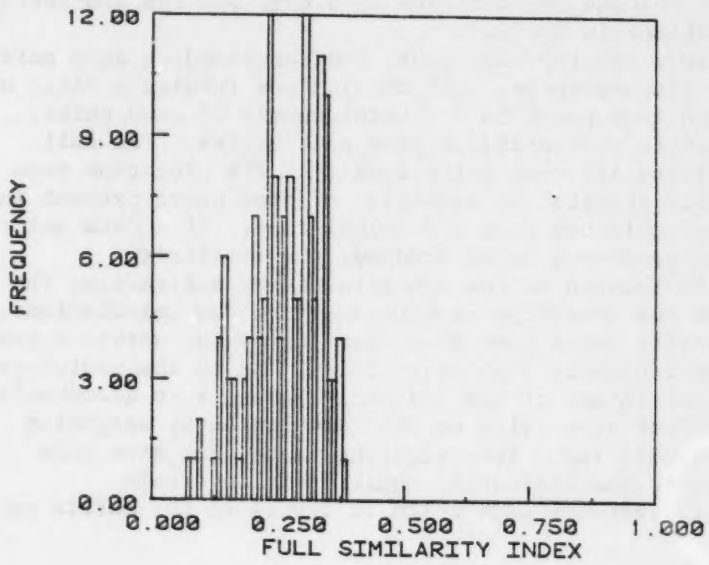

Figure 1. NNO Full Similarity Distribution

distribution is 0.261 and the standard deviation is 0.071. This distribution shows that there is some similarity within the data set but it is heterogeneous. The variance in the distribution may indicate the presence of several clusters of structurally related compounds in the data set. The several compounds with the lowest similarities might be considered as outliers from the rest of the data set. The four least similar compounds all have the N-nitroso group on an anilino nitrogen. A search using substructures indicates that only 11 of the 150 compounds have anilino nitroso groups and all 11 of these have similarities below the mean. In a SAR study special attention would have to be given to ensure that the anilino nitroso groups were suitably described. This example demonstrates the usefulness of a similarity measure based on an atom pair description of a molecule as a tool for exploring a group of compounds.

A more homogenous data set would be expected to have a similarity distribution with a mean much closer to unity and with

significantly less spread. The 9-AA data set is an example of such a set. Its similarity distribution, as seen Figure 2, shows half the

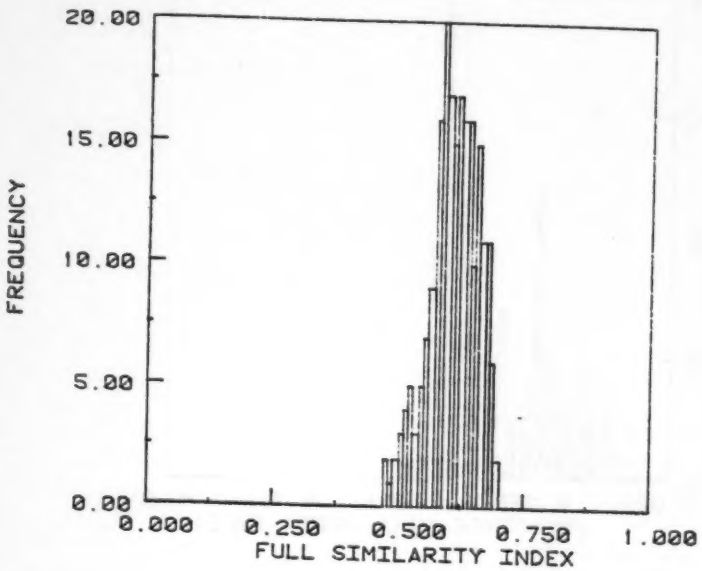

Figure 2. 9-AA Full Similarity Distribution

variance of the NNO data set and its mean of 0.607 is much closer to one. This is because all the 9-AAs share the same 21 atom 9-anilinoacridine backbone. The average similarity of the unsubstituted parent structure to the 213 members of the set is 0.561. The distribution plot does not indicate the presence of any structures that might be considered outliers. An average similarity distribution for the 9-AA data set was computed using the prototype index and is shown in Figure 3. The unsubstituted backbone was used as the prototype structure. The mean for the distribution is 0.351 and the standard deviation is 0.100. This mean is approximately half that of the full similarity distribution, and the prototype distribution spread is twice that of the full similarity distribution. The prototype similarity index is successful in emphasizing the diversity of substituents of the unsubstituted parent structure. The prototype distribution shows several compounds with small similarities. The least similar structure based on the prototype metric is the parent structure itself. With the subtraction of the atom pairs attributed to the 9-anilinoacridine structure, the parent can have no atom pairs in common with any of the remaining 212 training set compounds. The next six compounds in order of least similarity to the entire set all have a single simple or unique substituent on the anilino ring. The six compounds that had the least similarity in the full similarity distribution all have a single large substituent at the 1' position of the anilino ring. For data sets that share a common backbone, the prototype index is

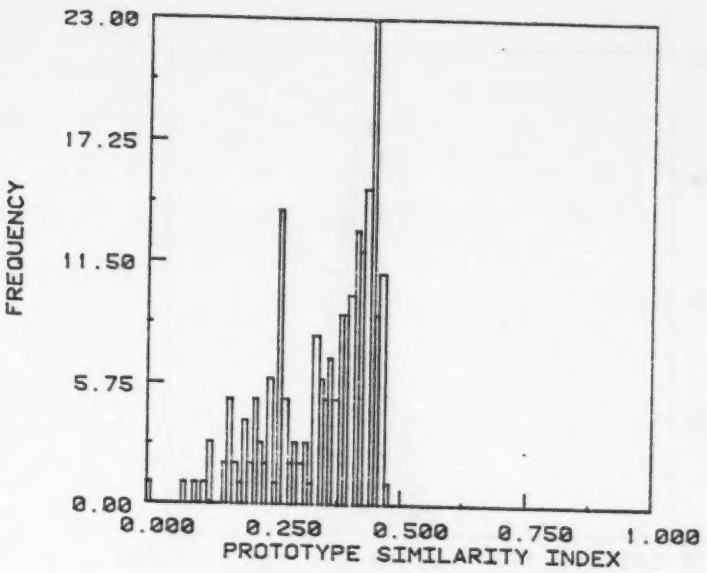

Figure 3. 9-AA Prototype Similarity Distribution

useful in identifying the compounds in the set with small different substituents while the full similarity index is useful in distinguishing those compounds with large distinctive substituents.

The full similarity distribution for the 92 NNO prediction compounds is shown in Figure 4. The similarity for each prediction set member was averaged over the entire training set of 150 NNO compounds to which each was compared. The mean and standard deviation of the prediction distribution is comparable to those of the training set distribution. However, because these are prediction compounds, it is not the position of each relative to the others that is as important as is the similarity relative to the distribution of the training set. The prediction distribution appears to be bimodal. Seven of the eight least similar compounds have nitroso groups attached to primary nitrogens. All of the training set compounds have nitroso groups on secondary nitrogens. The appropriateness of predicting compounds that have primary N-nitroso groups with a model that was developed entirely with secondary N-nitroso compounds should be questioned but will be determined by the extent to which the descriptors used in the model are unique for secondary N-nitroso functional groups. The least similar prediction compound is the N-nitroso derivative of folic acid. It has 34 atoms which is significantly above the mean of 10 for the training set. Its atom pair description consists of 319 unique atom pairs and a total of 536 pairs. The corresponding values for the training set are an average of 33.4 unique atom pairs per structure and an average of 56.0 total atom pairs per structure. With 536 total atom pairs it is not

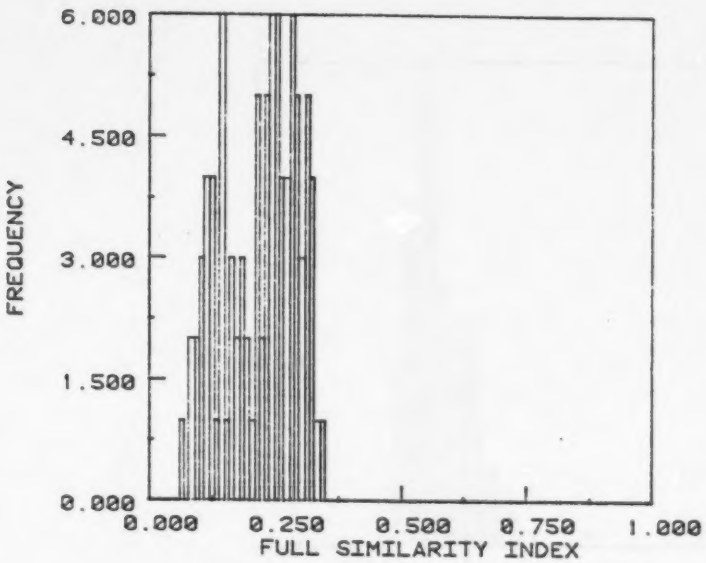

Figure 4. Prediction NNO Similarity Distribution

difficult to understand why the average similarity index for N-nitrosofolic acid is 0.069. This is an example where a model constructed of descriptors encoding just a simple moiety might not be sufficient to describe the activity of the entire molecule. The atom pair similarity metric proved useful in discovering this fact, and the case illustrates the need for whole molecule descriptions.

The PAC data set consists of 300 compounds ranging in size from benzene to seven ring systems. The fused aromatic systems provide a relatively small number of atom pair types. The frequency of the various unique atom pairs is related to the number of rings in the system. The full similarity distribution for the PAC data set is shown in Figure 5. The mean for the distribution is 0.629 and the standard deviation is 0.066. One would not expect the mean for the PAC distribution to be larger than that for the 9-AA distribution. The explanation can be found in the number of unique atom pairs found for each data set. The 9-AA data set has 1531 unique atom pairs as compared to 596 for the PAC set and 512 for the NNO set. The 65519 total atom pairs in the PAC set are thus distributed over a much smaller set of unique atom pairs than are the 29901 total atom pairs of the 9-AA set. The PAC distribution shows a larger spread than does the 9-AA distribution. This is due predominantly to the range of molecular sizes. Several potential outliers can be observed from the distribution. The three least similar compounds in the set are benzene, 1-naphthol, and fluorene, in order of increasing similarity. Given that these are so much smaller than the rest of the three- to seven-ring systems in the set, it might be appropriate to discard

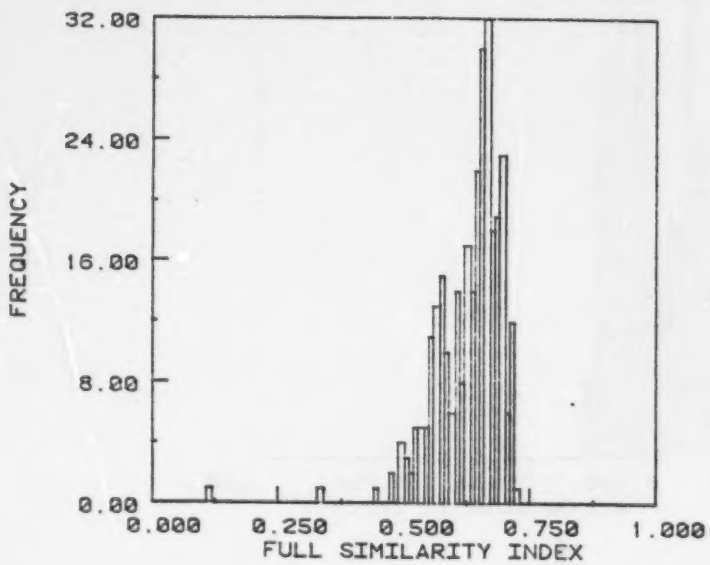

Figure 5. PAH Full Similarity Distribution

them from the set. As with the 9-AA set, the full similarity index is useful in pointing out those compounds in the PAC set with large distinctive substituents. The larger the substituents are relative to the parent structure, the larger the number of unique atom pairs generated will be and the lower the similarity. A similarity distribution for the PAC set based on the frequency similarity index was also generated, though not shown here. As with the prototype similarity index, the frequency similarity index is able to distinguish those compounds in the set with small but unique substituents or moieties. In the PAC set, the seven least similar compounds, as judged by the frequency similarity index, have unique substituent groups: -SH, -CN, -N=S and heteroatomic bridges between aromatic rings.

CONCLUSIONS

The atom pair method is useful as a whole molecule description based on topological information. The various similarity indices based on atom pairs serve as convenient tools for exploration of the extent and specifics of heterogeneity within a set of compounds. The selection of appropriate compounds for a prediction set for testing a model or a discriminant can be aided using the atom pair similarity measures.

ACKNOWLEDGEMENTS

This work was supported by Cooperative Agreement-Number CR 807531 from the U.S. Environmental Protection Agency. The contents do not necessarily reflect the views of the U.S.E.P.A. and no official endorsement should be inferred. The PRIME 750 computer was acquired with partial financial support from the National Science Foundation.

REFERENCES

1. R. E. Carhart, D. H. Smith, R. Venkataraghavan, Jour. Chem. Inf. Comp. Sci., 25 (1985), 64.
2. S. L. Rose and P. C. Jurs, Jour. Med. Chem., 25 (1982), 769.
3. W. Lijinsky, in T. K. Rao, W. Lijinsky, J. L. Epler (eds.), Genotoxicology of N-Nitroso Compounds, Plenum Pub. Co., New York, 1984, Chap. 10.
4. R. Preussmann and B. W. Stewart, in C. E. Searle (ed.), Chemical Carcinogens, 2nd Ed., American Chemical Society, Washington, DC, 1984, Chap. 12.
5. D. R. Henry, P. C. Jurs, W. A. Denny, Jour. Med. Chem., 25 (1982), 899.
6. E. K. Whalen-Pedersen and P. C. Jurs, in M. Cooke and A. J. Dennis (ed.), Chemical Analysis and Biological Fate: Polynuclear Aromatic Hydrocarbons, Battelle Press, Columbus, Ohio, 1981, pp. 55-61.
7. A. J. Stuper, W. E. Brugger, P. C. Jurs, "Computer Assisted Studies of Chemical Structure and Biological Function," Wiley-Interscience, New York, 1979.